A Comparative Analysis of Data Mining Techniques on Breast Cancer Diagnosis Data using WEKA Toolbox

Majdah Alshammari[1], Mohammad Mezher[2]
Department of Computer Science
Fahad Bin Sultan University, Tabuk, KSA

Abstract*—Breast cancer is considered the second most common cancer in women compared to all other cancers. It is fatal in less than half of all cases and is the main cause of mortality in women. It accounts for 16% of all cancer mortalities worldwide. Early diagnosis of breast cancer increases the chance of recovery. Data mining techniques can be utilized in the early diagnosis of breast cancer. In this paper, an academic experimental breast cancer dataset is used to perform a data mining practical experiment using the Waikato Environment for Knowledge Analysis (WEKA) tool. The WEKA Java application represents a rich resource for conducting performance metrics during the execution of experiments. Pre-processing and feature extraction are used to optimize the data. The classification process used in this study was summarized through thirteen experiments. Additionally, 10 experiments using various different classification algorithms were conducted. The introduced algorithms were: Naïve Bayes, Logistic Regression, Lazy IBK (Instance-Bases learning with parameter K), Lazy Kstar, Lazy Locally Weighted Learner, Rules ZeroR, Decision Stump, Decision Trees J48, Random Forest and Random Trees. The process of producing a predictive model was automated with the use of classification accuracy. Further, several experiments on classification of Wisconsin Diagnostic Breast Cancer and Wisconsin Breast Cancer, were conducted to compare the success rates of the different methods. Results conclude that Lazy IBK classifier k-NN can achieve 98% accuracy among other classifiers. The main advantages of the study were the compactness of using 13 different data mining models and 10 different performance measurements, and plotting figures of classifications errors.

Keywords*—Data mining; breast cancer; data mining techniques; classification; WEKA toolbox

I. Introduction

Worldwide, breast cancer has become one of the most common cancers [1]. It originates in the area of the breast tissue that has a concentration of milk ducts. Although most cases occur in women, there have been reported cases in men as well. There are noticeable signs and symptoms of breast cancer. The first noticeable symptom is usually a different mass from the rest of the breast tissue. Most women, about 80%, discover these masses during self-examinations.

Breast cancer can be classified as benign or malignant; however, this classification is determined through diagnostic testing. Some criteria to consider are uniformity of cell size and shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, and normal nucleoli. By observing these criteria, doctors or scientists are able to make a diagnosis according to the patient's diagnostic test results.

Certain unhealthy lifestyle choices tend to put some people in danger of developing breast cancer: smoking, consuming fatty food and alcohol, lack of exercise, and stress. Although genetics plays a minor role in many breast cancer cases, unhealthy habits contribute more.

The rapid spread of breast cancer and the inability to accurately diagnose and recognize its presence represents a challenge for researchers and developers in biomedical engineering [2]. This challenge leads to deploying new data mining techniques. Data mining is the uprooting and recall of unknown data from the past that can be useful. Data mining also includes the acknowledge recovery and analysis of data that is saved in a data repository. Some of the important methods of data mining are classification, association, clustering and regression, etc. [3]. The focus of this paper is to drive the research towards new feasible solutions for mining breast cancer data. Thus, a data mining-based experiment for breast cancer classification mechanism is introduced with different types of classifiers. In addition to identifying the best classifier model that introduces higher classification accuracy for the predefined dataset used in this study, the data mining process is implemented by applying pre-processing operations and extracting features to the specified data records from the data set using WEKA.

The WEKA (Waikato Environment for Knowledge Analysis) is an open-source software that contains a set of algorithms for data mining tasks [4]. These algorithms can be applied to a data set either directly through the WEKA interface or via Java code. Then the different classifiers are implemented with different variables using several algorithms and multiple options to compute the best accuracy ratio.

In this study, the classification process was summarized through 13 experiments, including three experiments using the Bayes Net algorithm by three different search mechanisms and ten experiments using classification algorithms, Naïve Bayes (NB), Logistic Regression, Lazy IBK (Instance-Bases learning with parameter K), Lazy Kstar, Lazy Locally Weighted Learner (LWL), Rules ZeroR, Decision Stump, Decision Trees J48, Random Forest, and Random Trees to create a predictive

model that can be tested with new records and that can obtain classification accuracy, and compare the results obtained after implementing different algorithms compared to the slow algorithm IBK and k-NN. However, k-NN was the best in ranking for optimum time and accuracy values. The optimal classifier was determined to identify more new records for accurate breast cancer identification.

Moreover, this study aimed at utilizing data mining techniques to diagnose breast cancer using diabetic patients' datasets [5]. By looking at the literature, it is noticeable that there have been many efforts to use data mining for breast cancer datasets; however, previous studies lack in comparing WEKA with different parametric values and attributes. Experiments in this research used thirteen different data mining algorithms as well as the use of feature selection for data cleansing. This study shows competitive results compared to previous studies, mentioned in the literature.

The organization of this paper is as follows: Related works on breast cancer datasets using thirteen different classifiers are discussed in Section II, followed by Section III which provides an introduction about the classification algorithm used in the experiment. Section IV introduces the methodology used in this work; all the data mining techniques which are compared and analyzed are illustrated in Section V. Section VI concludes the proposed study and highlights the most accurate classifiers.

II. Related Work

The comparisons made in [6], were based on the performance of four different machine learning algorithms: Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (C4.5) and k-Nearest Neighbors (k-NN) and were conducted on the Wisconsin Breast Cancer (WBC) datasets. The objective of the study and the experiments that were conducted were to determine the effectiveness of each algorithm in regards to precision, accuracy, specificity, and sensitivity. The results showed that SVM obtained 97.13% accuracy and outperformed Naïve Bayes, C4.5, and k-Nearest Neighbors algorithm (k-NN) that obtained accuracy variance between 95.12% and 95.28%.

In [7], Genetic Algorithm (GA) was used alongside different data mining techniques for WBC. GA was used to extract significant and informative features to reduce computational complexity and enhance the data mining processing speed. Data mining techniques used in this study were: Decision Trees (DT), Bayesian Network (BN), Logistic Regression (LR), Random Forest (RF), SVM, Rotation Forest, Radial Basis Function Networks (RBFN) and Multilayer Perceptron (MLP). Two WBC medical datasets (WBC and Wisconsin Diagnostic Breast Cancer (WDBC)) were used to test the performance of the algorithm models. The highest accuracy of 99.48% was obtained by Random Forest and GA feature selection.

The study conducted by [8] aimed at diagnosing breast cancer using three different techniques, namely: SVM, DT, and Artificial Neural Network (ANN). The study was applied on the WDBC dataset from the UCI. Feature selection was applied to increase the effectiveness of the methods. The ensemble method yielded the best results among the used methods. It gave 98.77% accuracy, 98.05% sensitivity and 100% specificity.

In [9], the authors used three well-known data mining methods, namely, Naïve Bayes (NB), J48, and RBF Network, to develop prediction models for the survivability of breast cancer. The data, that contains 683 instances, was acquired from the UCI [5]. To develop the prediction models, data selection, pre-processing, and transformation were applied. The results obtained from the experiment showed that the Naïve Bayes performed the best with a classification accuracy of 97.36%, RBF Network resulted in a classification accuracy of 96.77%, and the J48 resulted in classification accuracy of 93.41%.

The work by [10] used 12 different machine learning techniques for the diagnosis of breast cancer. The techniques that were used are namely; NB, Decision Table, Ada Boost M1, J48, J-Rip, Logistics Regression, Lazy IBK, Lazy K-star, Multiclass Classifier, Multilayer–Perceptron, RF, and RT. WBCD dataset was used to train the model. Most of the applied methods scored above 94%. Only NB underperformed, compared to the other models, with an accuracy of 73.21%. RT and Lazy classifier algorithms outperformed the others with an accuracy close to 99%.

In [11], evaluated six different data mining techniques, namely: SVM, Bayes Network (BN), ANN, k-NN, Decision Tree (C4.5) and Logistic Regression. The WEKA tool was used for the experiment on the WBC dataset. SVM and BN yielded the highest accuracy of 97.28%. However, the BN classifier produced minimal time compared to SVM, which makes the BN classifier better.

In [12], researchers employed eight different data mining techniques for breast cancer prediction. The dataset used for the experiment was WPBC [5]. The experiments were done on four classification algorithms: SVM, DT C5.0, NB and k-NN and on four clustering algorithms: EM, K means, PAM and Fuzzy c-means. The experiments were implemented using the R programming tool. The results showed that classification algorithms have better performance than the clustering where SVM and DT (C5.0) had the best accuracy of 81% and Fuzzy c-means resulted in the lowest accuracy of 37%, among the tested algorithms. The study conducted by [13] utilized three data mining techniques to classify breast cancer as either malignant or benign. The techniques conducted on the WDBC breast cancer dataset [5] are, namely: LR, NB and DT. Results showed that Logistic Regression (LR) got the highest classification accuracy of 97.90% among the other two tested classifiers.

The study conducted by [14] proposed nested ensemble methods to distinguish benign tumors from malignant breast cancers. Each ensemble method contains "Classifiers", as well as "Metaclassifiers" that can have more than two classification algorithms. Metaclassifiers were developed in the two-layer nested ensemble. The dataset used for the experiments was WBDC. The proposed method (used by [14]) was compared to the conventional single classifiers such as BN and NB. The results indicated that the two-layer nested ensemble method outperforms the single classifiers.

To analyze breast cancer data, [15] utilized four different DT classification algorithms, namely, Classification and Regression Trees (CART), J48, Best First Tree (BF Tree) and DT (AD Tree). The experiment employed the WEKA tool, and the results demonstrated that the J48 classifier reached the highest accuracy of 99% whereas the CART algorithms resulted in 96% accuracy; AD Tree algorithm resulted in 97%, and BF Tree algorithm resulted in 98%.

For the experiment in this research study, k-NN achieved the highest accuracy of 98% whereas in [7,4,1], RF achieved the highest accuracy. The highest accuracy in [8] was achieved by the ensemble methods of SVM, DT, and ANN. By analyzing the literature, it is noticeable that different techniques got the highest accuracy in each study as follows: in the study conducted by [11], the highest accuracy was achieved by Bayes Network (BN), and SVM and DT (C5.0) got the best accuracy in [12]. Similarly, by looking at [13], it is noticeable that the highest accuracy was yielded by LR and in [15] the highest accuracy was achieved by the J48 classifier.

By analyzing the literature, it was also noticed that the proposed research yields competitive results in terms of accuracy. However, not all mentioned previous works that used WEKA tools for data mining, the data mining using WEKA tools, achieved the same task. For example, [10], [13] and [14] used data mining methods to classify cases of breast cancer into malignant and benign. Moreover, the other studies did experiments on a few techniques while this study tested thirteen different algorithms.

Table I shows all the previous studies where a different methodology for utilizing data mining techniques to diagnose was used.

TABLE I. SHOWS COMPARISONS BETWEEN PREVIOUS STUDIES

Ref.	Model	Dataset	Highest performance	Accuracy result
[6]	NB, SVM, C4.5, k-NN	WBC	SVM	97.13%
[7]	GA, DT, BN, LR, RF, SVM, RF, RBFN, MLP	WBC, WDBC	RF	99.48%
[8]	SVM, DT, ANN	WDBC	SVM, DT, ANN	98.05 %
[9]	NB, J48, RBF Network	WDBC	NB	97.36%
[10]	NB, Decision Table, Ada Boost M1, J48, J-Rip (LR), Lazy IBK, Lazy K-star, NN, RF, RT	WDBC	Lazy IBK, Lazy K-star, RF, RT	99.14%
[11]	SVM, BN, ANN, k-NN, DT(C4.5) and LR	WBC	SVM, BN	97.28%
[12]	SVM, Decision Trees C5.0, NB, k-NN, EM, K means, PAM, Fuzzy c-means	WPBC	SVM, DT (C5. 0)	81%
[13]	LR, NB, DT	WPBC	LR	97.90%
[14]	**BayesNet, NB**	**WDBC**	**BN**	**98.07%**
[15]	CART, J48, BF Tree, AD Tree	Congressional Voting Records Data Set	J48 classifier	99%

III. CLASSIFICATION ALGORITHM

The experiment for this study ran the k-NN algorithm on the dataset. However, this algorithm is known as IBK in WEKA toolbox. The IBK classification algorithm, for each test instance, measured the distance to identify the nearest k instances to that instance from the training data. Then all selected instances were used for explication of prediction results. Such mechanism is referred to as k-NN algorithm. Fig. 1 shows the pseudo code of the k-NN algorithm that is described as follows [16,17]:

Consider (Xi, Ci) where i = 1, 2…, n are the points of data. Xi is the feature values & Ci are the labels for Xi for each i. Considering 'c' is the number of classes then for all values of i, classes are represented as Ci ∈ {1, 2, 3, ……, c}.

Consider x is a point of an unknown label, and finding that unknown label class using k-NN algorithms is performed by:

Step 1: D (x, xi) = ECULIIAN (x, xi), i = 1,2, 3,…n
Step 2: D_s = SORT (D (x, xi)) in Ascending
Step 3: GET k elements from top of D_s, K is +int
Step 4: Get Indices of the k elements
Step 5: while k≥0, ki refers to k points belongs to i_{th} class
Step 6: If ki >kj ∀ i ≠ j then put x in class i

Fig. 1. Pseudo-Code of the IBK Algorithm.

In the IBK algorithm, a model was not manufactured but generated a just-in-time prediction for a test case. In each instance, the IBK algorithm reached a distance measure for locating k "near" cases in training data and used those instances to predict in order to determine which classifier in the WEKA toolbox, using the diabetic patient's dataset, had the highest accuracy. Experiments in this research analyzed the comparison of thirteen algorithms in various accuracy measurements.

Each algorithm conducted is represented briefly in terms of how it operates with the key parameters of the respective algorithm. The parameters of the algorithms conducted in this study are highlighted in Table III. The region size of k-NN is verified by the k-parameter. The distance metric used in k-NN is another important parameter. In the k-NN algorithm, the distance metric in default is Euclidean distance, which is ideal for quantitative data of the same size to determine the distance between instances.

The parameter C called the complexity parameter in WEKA governs the versatility with which the line can be drawn to separate groups. There is no margin violation with a value of 0, while the default is 1. The type of kernel to be used is a key parameter in SVM. The easiest kernel is a linear kernel that separates data from a direct line or hyperplane. The default kernels for the WEKA is the polynomial, where the higher the polynomial kernel, the wigglier of the exponent value. The classes are separated by means of a curved or angled line. In LR algorithm, the ridge parameter determines how much the algorithm needs to be forced to decrease the coefficient value. This regularization is deactivated by setting it to 0.

In order to implement the NB algorithm, the Kernel Estimator argument was used which might more accurately suit the actual distribution of attributes in the dataset. With DT, the researchers of this study chose to transform the no Pruning parameter to Real value. The minimum number of the tree instances in a leaf node when constructing the tree out of the training data was set to 7.

IV. Methodology

The methodology followed in this paper first started with data collection and data preparation from the BC dataset. Then data mining techniques were applied to generate a classification model that was used for evaluation and deployment. The diagrammatic representation of the proposed study framework is shown in Fig. 2. This methodology framework represents the standard data mining framework that should be followed in various applications of data mining, and the framework states that before data processing, selection and pre-processing should be performed to get clean and complete data records for processing. Then transformation of data represents a data extension conversion or compatibility conversion of data shape to an equivalent data mining tool format to perform data mining algorithm implementation and generate the desired knowledge outcome.

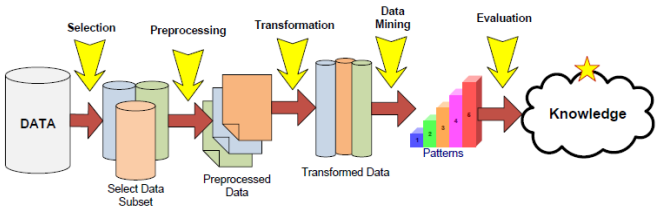

Fig. 2. Study Framework.

Analysis of the breast cancer dataset using WEKA machine learning software tool aims for mining the relationship in breast cancer data for efficient classification. A dataset is an aggregation of data which refers to the contents of one statistical data matrix or one database table. This dataset is then processed using WEKA toolbox. In this sense, Table I provides an overview of the dataset used in the experiments. The dataset values for each of the variables, such as the height and weight of an object, are then listed and each member of the dataset is indicated by a datum [17-18]. The dataset may include data for single or multiple members, with respect to the number of rows [19-20].

For this study, diabetic patients' dataset was collected and consists of 699 patients records with 10 different attributes and nine nominal attributes of them were selected as shown in Table II. The data mining algorithms were explored to identify efficient classification of diabetes dataset [5]. Accuracy metric was used as the main comparison base while other metrics were also considered such as the Precision Recall (PRC), corresponding sensitivity (recall), the Receiver Operating Characteristics (ROC) and Matthews Correlation Coefficient (MCC).

TABLE II. Dataset Attributes Description

Description	Invalid records	Range	Mean Dev.	Standard Dev.
Clump thickness	45	10-1	4.442	2.821
Uniformity of cell size	35	10-1	3.151	3.065
Uniformity of cell shape	-	10-1	3.215	2.989
Marginal adhesion	70	10-1	2.83	2.865
Single epithelial cell size	-	10-1	3.234	2.223
Bare nuclei	30	10-1	3.545	3.644
Bland chromatin	-	10-1	3.445	2.45
Normal nucleoli	25	10-1	2.87	3.053
Mitoses	-	10-1	1.603	1.733

V. Comparison with other Algorithms

In this section, an experimental analysis of the effectiveness of the proposed methodologies for the thirteen different classifiers using the same dataset was completed. The experiments were implemented using WEKA toolbox. The experiments were done on a Toshiba desktop computer with Intel(R) Core (TM)i7-4710MQ CPU with @2.50GHz, 2.5GHz, and 8192MB RAM. The result of the evaluation represents the classification model results used to evaluate 25% of the selected dataset records, while the remaining 75% was used for training.

Table III shows the different algorithms used in the experiments' setup. The experimental methodology of the study, used 11 different classification algorithms and 3 other search methods.

TABLE III. Algorithms used in the Experiments' Setup

No.	Algorithm Variant
1	BayesNet - Tabu Search
2	BayesNet - K2 Search
3	BayesNet - TAN Search
4	Naive Bayes
5	Logistic
6	Lazy – IBK
7	Lazy – Kstar
8	Lazy – LWL
9	Rules ZeroR
10	Trees DecisionStump
11	Trees J48
12	Trees RandomForest
13	Trees RandomTree

Fig. 3 represents the execution of the performance records registered by WEKA. The figure shows evaluation of classification for the PRC, ROC area and MCC results. Where Precision Recall (PR) is the precision values for corresponding sensitivity (recall) values and the ROC area refers to general performance of the classifier. The MCC is a measure of the quality of the binary classifications. However, the general combined model measures the precision and recall calculated in the F-Measure as in (1).

$$F = 2 * Precision * Recall / (Precision + Recall) \quad (1)$$

So, alternative performance measurement was the confusion matrix specifically using the ROC curve shown in Fig. 3. However, the accuracy of the Algorithm 9 (Rules-ZeroR algorithm) was very weak; in contrast, PRC is more useful which shows how the classifier was behaving on one class.

The classifiers from strong and weak classifications (Fig. 3) were reviewed and analysed in order to investigate these results further (Table IV). Groups of classifiers that were very similar in their performances were found. Thereby, similar performance from the similar classifiers are highlighted in bold in Table IV.

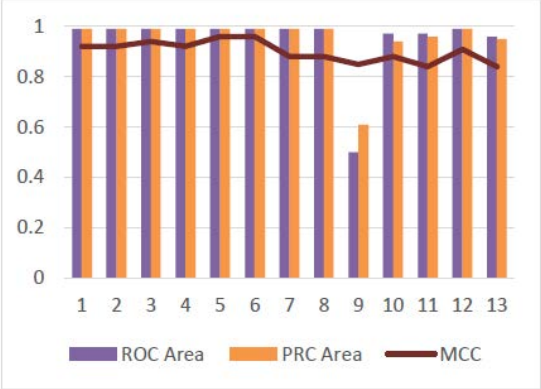

Fig. 3. Performance Results for All Experiments.

TABLE IV. EXPERIMENTS INDICES USED IN EVALUATION

No	F-Measure	Recall	Precision
1	0.97	0.97	0.97
2	0.97	0.97	0.97
3	0.98	0.98	0.98
4	0.97	0.97	0.97
5	**0.99**	**0.99**	**0.99**
6	**0.99**	**0.99**	**0.99**
7	0.95	0.95	0.96
8	0.95	0.95	0.96
9	0.85	0.73	0.73
10	0.95	0.95	0.96
11	0.94	0.94	0.94
12	0.97	0.97	0.97
13	0.94	0.94	0.94

In Fig. 4, Roles ZeroR had the weakest classifier which ultimately lead to a lower accuracy rate while other investigated algorithms produced more than 90% accuracy rate. Fig. 4 also shows that the best performed classifier in the three measurements was for both Algorithm 5 (logistic algorithm) and Algorithm 6 (lazy-IBK algorithm). In Fig. 4, the thirteen classifiers' results show both the weak (shown in orange color) and strong classifiers' (shown in blue color) performance. The classifiers were experimented for the dataset illustrated in Table II.

The statistical records for the execution of the experiments shown in Fig. 5, are the kappa statistics, absolute error, and the mean error. The failure classifier in the experiment was for Algorithm 5 (logistic algorithm). All statistics correspond to the classifiers' results where the type of errors is computed as a part of Kappa statistics plots.

Fig. 5 shows the final results of all experiments which are introduced in Table V. Table V shows the accuracy and elapsed time during the setup experiment time of each algorithm and its configurations. The worst performance was by the Rules ZeroR algorithm, and the best by the Lazy IBK and K-star algorithms while all other tests introduced more than 90% accuracy. The highest performance results are grouped and highlighted in bold.

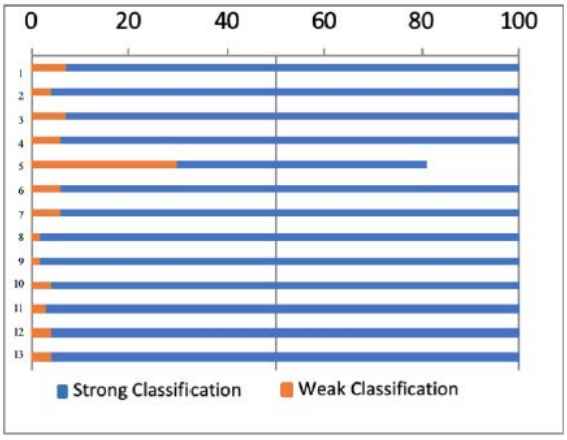

Fig. 4. Strong and Weak Classifications.

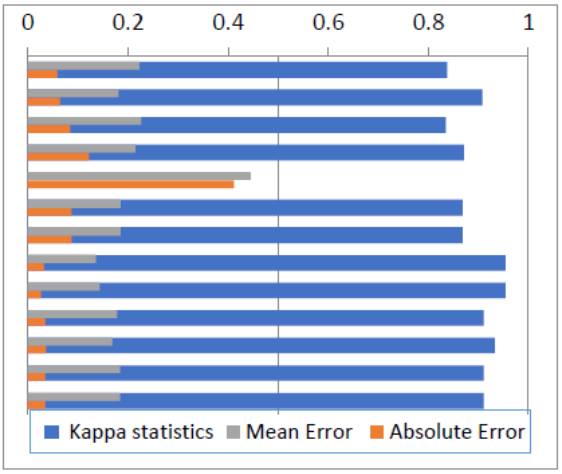

Fig. 5. Error and Kappa Statistics.

TABLE V. FINAL ACCURACY AND ELAPSED TIME RESULTS

No	Classifier	Accuracy	Elapsed Time
1	BayesNet- TabuSearch	96.4%	0.01
2	BayesNet- K2 search	96.4%	0
3	BayesNet TAN search	97.3%	0
4	Naive Bayes	96.4%	0
5	Logistic	92.8%	0.15
6	Lazy – IBK	**98.2%**	0
7	Lazy – Kstar	**98.2%**	0.92
8	Lazy – LWL	94.6%	0.17
9	Rules. ZeroR	73 %	0
10	Trees. DecisionStump	94.6%	0
11	Trees.J48	93.7%	0.02
12	Trees. RandomForest	96.4%	0.04
13	Trees. RandomTree	93.7%	0

VI. CONCLUSIONS

This paper studied a common practical problem in the detection or recognition of data patterns using data mining techniques. The comparative analysis proposed here for the Breast Cancer dataset using different pre-processing techniques was conducted using the WEKA data mining tool. The final evaluation of classification processes was done by extracting accuracy ratios for all experiments, and the results showed high rates ranging between 72% and 98%. The other ten experiments used the different classification algorithms to obtain the highest accuracy ratios; however, the IBK and K-star algorithms from Lazy algorithms showed the best performance up to 98.2% in optimum time and accuracy values. Those records can be further used in real-world applications such as any development models introduced for biomedical laboratory technology.

In the future, this study will be extended by utilizing deep learning techniques in order to get the highest accuracy. Moreover, the proposed technique will be applied on datasets for different cancer types.

REFERENCES

[1] Otoom, Ahmed Fawzi, Emad E. Abdallah, and Maen Hammad. Breast Cancer Classification: Comparative Performance Analysis of Image Shape-Based Features and Microarray Gene Expression Data. International Journal of Bio-Science & Bio-Technology, Vol.7, No.2 ISSN: 2233-7849, pp.37 – 46, 2015.

[2] A. Thomas, A. Rhoads, E. Pinkerton, M. C. Schroeder, K. M. Conway, W. G. Hundley, et al., Incidence and Survival Among Young Women with Stage I-III Breast Cancer: SEER 2000-2015. JNCI cancer spectrum, vol. 3, pp. pkz040-pkz040, 2019.

[3] Sayedeh Somayeh Hosaini and Mehran Emadi Breast Cancer Tumor Diagnosis from Mammography Images Using Wavelet Transform and Hidden Markov Model. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 4, Issue 8, ISSN: 2278 – 8875, pp. 6815-6823, 2015.

[4] Hajar Saoud, Abderrahim Ghadi, Mohamed Ghailani, Boudhir Abdelhakim.2018. Application of Data Mining Classification Algorithms for Breast Cancer Diagnosis: SCA '18: 3rd International Conference on Smart City Applications Tetouan Morocco October, 2018 ISBN:978-1-4503-6562-8.

[5] Breast Cancer Wisconsin Data Set UCI. 1992.

[6] Chaurasia, V., Pal, S. & Tiwari, B. B. 2018. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms and Computational Technology 12(2): 119–126. doi:10.1177/1748301818756225.

[7] Aličković, E. & Subasi, A. 2017. Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and Applications 28(4): 753–763. doi:10.1007/s00521-015-2103-9.

[8] Zorluoglu, G. & Agaoglu, M. 2017. Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods 2 Related Works 2 Classification Methods The classification models of Clementine used in 2: 24–27.

[9] Chaurasia, V., Pal, S. & Tiwari, B. B. 2018. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms and Computational Technology 12(2): 119–126. doi:10.1177/1748301818756225.

[10] Kumar, V., Mishra, B. K., Mazzara, M., Thanh, D. N. H. & Verma, A. 2020. Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. Lecture Notes on Data Engineering and Communications Technologies, hlm. Vol. 37. Springer Singapore. doi:10.1007/978-981-15-0978-0_43.

[11] Saoud, H., Ghadi, A., Ghailani, M. & Boudhir, A. A. 2018. Application of data mining classification algorithms for breast cancer diagnosis. ACM International Conference Proceeding Series. doi:10.1145/3286606.3286861.

[12] Ojha, U. & Goel, S. 2017. A study on prediction of breast cancer recurrence using data mining techniques. Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data ScienceandEngineering527–530. doi:10.1109/CONFLUENCE.2017.7943207.

[13] Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree. International Journal Of Engineering And Computer Science 6(2): 20388–20391. doi:10.18535/ijecs/v6i2.40.

[14] Barua, P. D. & Gururajan, R. 2020. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters 132: 123–131. doi: 10.1016/j.patrec.2018.11.004.

[15] Rodríguez-Jiménez, J. M., Cordero, P., Enciso, M. & Mora, A. 2016. Data mining algorithms to compute mixed concepts with negative attributes: an application to breast cancer data analysis. Mathematical Methods in the Applied Sciences 39(16): 4829–4845. doi:10.1002/mma.3814.

[16] Ibrahim, Amal S., and Nabiel NH Mikhail. "The Evolution of Cancer Registration in Egypt: From proportions to population-based incidence rates." SECI Oncology 2015 DOI: 10.18056/seci, pp.1-21, 2015.

[17] Shraddha Soni. “A Literature Review on Data Mining and its Techniques “Indian journal of applied research, Volume 05- issue 06, ISSN: 2249-5555, pp.730-731, June. 2015.

[18] Ibrahim, Amal S., and Nabiel NH Mikhail. The Evolution of Cancer Registration in Egypt: From proportions to population-based incidence rates. SECI Oncology 2015 DOI: 10.18056/seci, pp.1-21, 2015.

[19] Fariba Mirbaha, Gloria Shalviri, Bahareh Yazdizadeh, Kheirollah Gholami and Reza Majdzadeh. Perceived barriers to reporting adverse drug events in hospitals: a qualitative study using theoretical domains framework approach. Implementation Science, 10: 110 DOI: 10.1186/s13012-015-0302-5, pp.1-10, 2015.

[20] Sayedeh Somayeh Hosaini and Mehran Emadi Breast Cancer Tumor Diagnosis from Mammography Images Using Wavelet Transform and Hidden Markov Model. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 4, Issue 8, ISSN: 2278 – 8875, pp. 6815-6823, 2015.